

CLUSTER ANALYSIS IN DEGRADED AREAS BY EROSION IN THE NAZARENO (MG) REGION

Raul Cassaro¹, Rogério R. M. Ferreira², Vinicius R. M. Ferreira³, Valéria G. S. Rodrigues¹

¹ São Carlos School of Engineering, University of São Paulo, São Carlos SP 13566-590, Brazil

² Brazilian Agricultural Research Corporation - 303 Soldado Passarinho Ave. - 13070-115- Campinas - SP, Brazil

³ Siriema Produtos Ambientais -Ltda., São João del Rei/MG, Brazil
valguima@usp.br

ABSTRACT

When analyzing a database, summarize collected information is the analyst main challenge. In many cases, when you have a large number of data, it may be of interest to create groups. A cluster analysis is a multivariate statistical procedure that tries to group a data set into homogeneous subgroups, called clusters. In this context, RStudio software has a large number of functions and work packages for cluster analysis. In the Nazareno region (MG), numerous studies have addressed the characterization of areas degraded by large erosive processes (gullies). Thus, this study aimed at a cluster analysis in a database composed of soils collected in 4 trenches and at different depths Nazareno's region (MG) gullies. The parameters obtained were: pH (pH in H₂O and KCl), granulometric fractions, organic matter, density, total pore volume, microporosity and macroporosity. An analysis of clusters partitioned the data into three groups: a) 1 group: trench 4 (depth from 0 to 40 cm), composed of more clayey soils and higher content of organic matter; b) 2 group: trenches 1 and 2 (depths 20-30 cm and 30-40 cm), composed of siltier soils and lower organic matter content; c) group 3: trenches 1, 2 and 3 (different depths). With a cluster analysis, it was possible to separate 3 groups, and these are more related to soil granulometry and organic matter, which are parameters of fundamental importance when analyzing areas degraded by erosion processes. With the results it becomes clear the places (trenches) and the depths with greater susceptibility to erosion.

Keywords: gully - multivariate analysis - Nazareno - Brazil.

RESUMEN

Al analizar una base de datos, resumir la información recopilada es el principal desafío del analista. En muchos casos, cuando tienes una gran cantidad de datos, puede ser de interés crear grupos. Un análisis de conglomerados es un procedimiento estadístico multivariado que intenta agrupar un conjunto de datos en subgrupos homogéneos, llamados conglomerados. En este contexto, el software RStudio tiene una gran cantidad de funciones y paquetes de trabajo para el análisis de conglomerados. En la región de Nazareno (MG), numerosos estudios han abordado la caracterización de áreas degradadas por grandes procesos erosivos (cárcavas). Así, este estudio tuvo como objetivo un análisis de conglomerados en una base de datos compuesta por suelos recolectados en 4 trincheras y a diferentes profundidades de los barrancos de la región de Nazareno (MG). Un análisis de conglomerados dividió los datos en tres grupos: a) 1 grupo: zanja 4 (profundidad de 0 a 40 cm), compuesta por suelos más arcillosos y mayor

contenido de materia orgánica; b) 2 grupo: trincheras 1 y 2 (profundidades 20-30 cm y 30-40 cm), compuestas por suelos más limosos y menor contenido de materia orgánica; c) grupo 3: zanjas 1, 2 y 3 (diferentes profundidades). Con un análisis de conglomerados se logró separar 3 grupos, y estos están más relacionados con la granulometría del suelo y la materia orgánica, que son parámetros de fundamental importancia a la hora de analizar áreas degradadas. Con los resultados se aclaran los lugares (trincheras) y las profundidades con mayor susceptibilidad a la erosión.

Palabras clave: cárcavas - análisis multivariable - Nazareno - Brasil.

Introduction

The Nazareno's city region (on Minas Gerais State, Brasil), has a history of anthropic degradation leading to severe a harm to the local soil. Since the region's gold exploration on XVII century, through the 1950's opening of dirt roads with no care about drainage works, until today with the lack of conservationist practices by the local along with the opening of ditches to mark the territorial division between properties (FERREIRA, 2005).

The local soil main characteristics are known already: Superficially, the soil alternates between red oxisol and red-yellow oxisol, both with high clay fraction, well structured and well developed with a deep B-horizon (more than 1 meter deep). Underneath it, there is a silty saprolite cambisol that is chemically poor, with no structure and easily erodible. This cambisol has a high concentration of kaolinite and aluminum, inhibiting the vegetation grown, decreasing the OM content and soil cohesion (FERREIRA, 2005; FERREIRA, 2008; SAMPAIO, 2014; OLIVEIRA, 2015; CASSARO, 2018).

This study aims to bring the relations between soil characteristics and erosion, using one of these previous studies on the region through the cluster analysis using R/RStudio program. R/RStudio is a free statistical software for analyzing, manipulating, calculating and visualizing data. It has some features like: easy access, handling and installation; a big and integrated toll collection; graphic resources for data exhibition; and an easy programming language. Between its statistical analyses are principal component methods, cluster analysis, and machine learning (KASSAMBARA, 2017).

Cluster analysis involves the distance calculation between data and its classification on different groups by the distance or (dis)similarity. The most commonly used unsupervised machine learning algorithm for portioning a given data set is the Hartigan-Wong (1979) k-means clustering, which defines the total within-cluster variation as the sum of squared distances Euclidean distances between items and the corresponding centroid (KASSAMBARA, 2017).

Materials and Methods

Data Source

Ferreira (2008) studied a dystrophic cambisol with a leucocratic gneiss-granite origin. The author collected soil inside four trenches (60 cm deep each one) and did soil analysis of

eight parameters: pH (pH in H₂O and KCl), granulometric fractions, organic matter, density, total pore volume, microporosity and macroporosity. Those parameters were obtained for four depths on the trenches: 0-10cm, 10-20cm, 20-40cm and 40-60cm. Those data were used to obtain the dissimilarity clusters.

Cluster Analysis

The cluster analysis was made using the R/RStudio software. At first, the data was scaled because the data needs to be comparable, since each parameter is in a different scale (ex: kilogram, kilometers, percentage, etc), so it needs to be standardized allowing the software to measure de dissimilarities between those parameters.

To choose the number of clusters were made tests of clustering with 2, 3, 4, 5 and 6 clusters. The only test that generated usable and with enough members was the one with 3 clusters (each one with 4, 4 and 8 members).

Results and discussion

The table with Ferreira (2008) data separated by clusters is on Figure 1

Leucocratic Gneiss-Granite Cambisol	pH (H ₂ O)	pH (KCl)	Clay (%)	CMS (%)	CS (%)	MS (%)	FS (%)	FMS (%)	OM (Dag/Kg)	Silt (%)	Density (g/cm ³)	TPV (%)	Micro (%)	Macro (%)
Trench 4 (0-10cm)	5	4	38	0	4	24	21	2	2,4	19	1,4	44,8	38,8	6
Trench 4 (10-20cm)	5	4	33	0	4	20	25	5	2,2	13	1,5	42,1	37,8	4,4
Trench 4 (20-40cm)	5,1	4	37	0	5	25	25	4	1,9	4	1,4	45,1	34,8	40,4
Trench 4 (40-60cm)	5,2	4,1	36	0	5	28	22	4	1,2	5	1,4	46,6	30,3	16,3
Trench 1 (20-40cm)	5,4	4,1	21	0	5	21	12	2	0,8	39	1,4	45,8	33,1	12,7
Trench 1 (40-60cm)	5,6	4,2	18	0	2	13	21	6	0,4	40	1,5	40,7	39,6	1,1
Trench 2 (40-60cm)	5,5	4,2	21	0	4	15	19	5	0,5	36	1,5	42,3	41,8	0,5
Trench 3 (40-60cm)	5,6	4,1	22	0	3	14	20	6	0,4	35	1,4	45,1	39	6,2
Trench 1 (0-10cm)	5,3	4,2	23	0	2	13	32	8	1,2	22	1,5	40,9	37,9	3
Trench 1 (10-20cm)	5,1	4,1	25	17	2	15	26	7	1,1	8	1,5	42,4	36,3	6,1
Trench 2 (0-10cm)	5,2	4,1	29	0	3	17	27	5	1,2	19	1,5	41,7	37,2	4,6
Trench 2 (10-20cm)	5,3	4,2	27	0	4	16	25	7	0,9	21	1,3	48,3	39,3	9,1
Trench 2 (20-40cm)	5,4	4,1	26	0	2	15	30	7	1,4	20	1,4	45,7	38,5	7,2
Trench 3 (0-10cm)	5,2	4,1	27	0	3	15	23	4	1,4	28	1,4	47,3	38,8	8,6
Trench 3 (10-20cm)	5,3	4,1	28	0	3	15	25	6	1,3	23	1,3	49,1	39,9	9,3
Trench 3 (20-40cm)	5,4	4,1	29	1	4	15	25	5	0,9	21	1,3	50	38,8	11,2

Fig.1. The table show the Ferreira (2008)'s results separated by the clusters. Each color (green, red and blue) represents one cluster. CMS – coarse medium sand; CS – coarse sand; MS – medium sand; FS – fine sand; FMS – fine medium sand; OM – organic matter.

The k-means algorithm created three groups (clusters). The first one (green) made a cluster with the clayey soils and with the highest organic matter content, all of them from the fourth trench. The second group (red) were created with the siltier soils and lowest OM

content, also three of the samples being the deepest of its respective trench. The third cluster (blue) contains the other soils without outstanding features.

The results were sufficient to show the cluster analysis capacity to aggregate important information. Only with the data, the clustering process was able to separate the clayey trench (green cluster) and the most susceptible to erosion trench (red cluster). The red cluster has the silty soil and the lowest OM content (all samples have low OM content, but here we are comparing between the samples), which has already been proven to be an important characteristic on a soil susceptible to erosion on the region (FERREIRA, 2005; FERREIRA, 2008; SAMPAIO, 2014; OLIVEIRA, 2015; CASSARO, 2018).

Conclusions

The findings were capable to show the potential of statistical tools like clustering when analyzing different scale parameters and trying to understand its relation. With the different clusters presented, it was possible to see the strong relation between the soil granulometry percentages. The first cluster has only clayey soils from the same trench and the second cluster has the silty soils and, not coincidentally, the deepest and the lowest OM percentage, since the silt high percentage was already noticed, by other authors, as a deep soils characteristic and a factor to induce the low OM presence, on the region. The third cluster with the other members, the ones without big variations in the average of the data collected. As the software was able to separate the soils between its main and more relevant characteristics (for erosion studies), it certainly has a potential to analyze different parameters for different purposes and establish relations not seen before.

Acknowledgements

The authors are grateful for the financial support by the Coordination for the Improvement of Higher Education Personnel (CAPES).

References

- CASSARO, R. Análise dos Processos Erosivos na Bacia do Córrego do Forro – Município de Conceição da Barra de Minas (MG): Estudo dos Condicionantes Geológicos. 2018. 156p. Dissertação (Mestrado em Ciências). Universidade de São Paulo, Escola de Engenharia de São Carlos, São Carlos. 2018.
- FERREIRA, R.R.M. Qualidade física de cambissolos sobre dois materiais de origem com pastagens extensivas. 2008. 106p. Tese (Doutorado em Agronomia). Universidade Estadual de Londrina, Londrina, 2008.
- FERREIRA, V.M. Voçorocas no município de Nazareno, MG: origem, uso da terra e atributos do solo. 2005. 84p. Dissertação (Mestrado em Agronomia). Universidade Federal de Lavras, Lavras, 2005.
- HARTIGAN, J.A. WONG, M.A. A K-means clustering algorithm. *Applied Statistics* 28, 100-108.
- KASSAMBARA, A. *Practical Guide To Cluster Analysis in R – Unsupervised Machine*